



Bionformatics tools to analyze Genomic Data

This course covers the state-of-the-art of major relevance in today's genomic data analysis. Through sessions of theory and practical examples, researchers will acquire the experience necessary to address scientific questions using genomic datasets and solve them.

Special attention will be devoted to important (although not always taken into account) aspects in high-throughput data analysis, such as multiple testing or functional annotation.

The course is designed to be a mixture of theoretical and practical sessions. The latter will require some familiarity with the use of web-based tools and knowledge of basic notions of statistics.

Practical sessions will be carried out using the **Babelomics suite** (Medina et al. 2010). We will also use others bioinformatics tools: CellBase (Bleda et al. 2012), FastQC and more resources to treat this kind of data.

Day 1

15.30 - 16.00. Introduction.

Why genomic data analysis? Pre- and post-genomics hypothesis testing: a note of caution. Design of experiments.

16.00 - 18.30. General overview about different bioinformatic tools.

Data preprocessing and normalization. Unsupervised analysis (clustering). Supervised analysis (gene selection, predictors). Functional annotation and profiling: FatiGO, Gene Set Analysis and SNOW.

18.30 - 19.30. Babelomics WEB usage introduction.

Short introduction to Babelomics Web environment and resources.

Day 2

15.30 - 16.30. Functional Annotation Databases.

Most popular gene and protein annotation repositories are revised. These are some of the data that the Babelomics suite imports for the functional profiling of genomic experiments.

16.30 - 17.00. Getting genes from the chromosome locations.

Converting genome coordinates. Retrieving biological information from chromosome locations: CellBase.

17.00 - 18.30. Single Enrichment Analysis: FatiGO.

Introduction to Single Enrichment Analysis methods. How to use FatiGO to explore your data in the view of most popular databases: Gene Ontology, KEGG Pathways, InterPro motifs...

18.30 - 19.30. Gene Set Enrichment Analysis: Fatiscan and Logistic Regression Analysis.

Introduction to Gene Set Analysis methods. Understanding the biological roles played by the genes in the experiments. Using different types of information for the functional profiling of high-throughput experiments: Gene Ontology, InterPro motifs, transcription factor binding sites, etc.

Day 3

15.30 - 16.30. SNOW: Studying Networks in the Omic World.

SNOW is a web-based tool that introduces protein-protein interaction data into the functional profiling of genome-scale experiments. It extracts from a list of pre-selected proteins or genes the minimal connected network (smallest network that connects all the elements of the list) that they conform in terms of physical interactions and then it evaluates its topological parameters comparing them versus same-size networks generated from random lists of genes/proteins.

16.30 - 18.00. Introduction to Next Generation Sequencing Technologies for Variations Studies.

NGS pipeline analysis: sequence preprocessing, mapping, variant calling, variant prioritization, functional annotation, GWAS analysis and Gene-Set analysis.

18.00 - 19.30. Quality Control for Next Generation Sequencing raw data.

Data formats: fasta and fastq. Sequence quality encoding. Quality control tools. FastQC.

References

Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tárraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, García F, Marbà M, Montaner D, Dopazo J. Nucleic Acids Res 2010. 38:W210-3

SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. Pablo Minguez¹, Stefan Götz^{1,2}, David Montaner¹, Fatima Al-Shahrour¹ and Joaquin Dopazo. Nucleic Acids Res 2009. 37:W109

Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. Al-Shahrour F, Carbonell J, Minguez P, Goetz S, Conesa A, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J. Nucleic Acids Res 2008. 36:W341-6

Babelomics: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J. Nucleic Acids Res 2005, 33:W460-464.

GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. Vaquerizas JM, Conde L, Yankilevich P, Cabezón A, Minguez P, Diaz-Uriarte R, Al-Shahrour F, Herrero J, Dopazo J. Nucleic Acids Res 2005, 33:W616-620.

CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Bleda M, Tárraga J, De Maria A, Salavert F, Garcia-Alonso L, Celma M, Martín A, Dopazo J, Medina I. Nucleic Acids Res 2012, 40:W609-14.